情報統計
第2回

2023年8月1日　　神奈川工科大学

櫻井 望
公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

統計の基本と用語

学習目標

以下の統計用語をマスターします

- 平均値、中央値
- 分散、標準偏差
- 統計量
- 分布
- 母集団
- ランダムサンプリング
- 標本
- 統計的推定
- 母平均、母分散
- 標本平均、標本分散、不偏標本分散
- 正規分布（ガウス分布）
- 標準誤差

復習

統計って？

集団の状況を
数値で表したもの

↑

目的：集団の〇〇を知りたい

復習

統計学

- データを集める
- 解析する
- 解釈する

ための**方法論**

↓

結果：集団の〇〇がわかった！

第1回の
身長データを使って
解析してみる

目的:
このクラスの人の身長は
どのくらい？

集めたデータ

集団の状況を表す
代表的な値を計算

平均値
中央値

中心を表す値

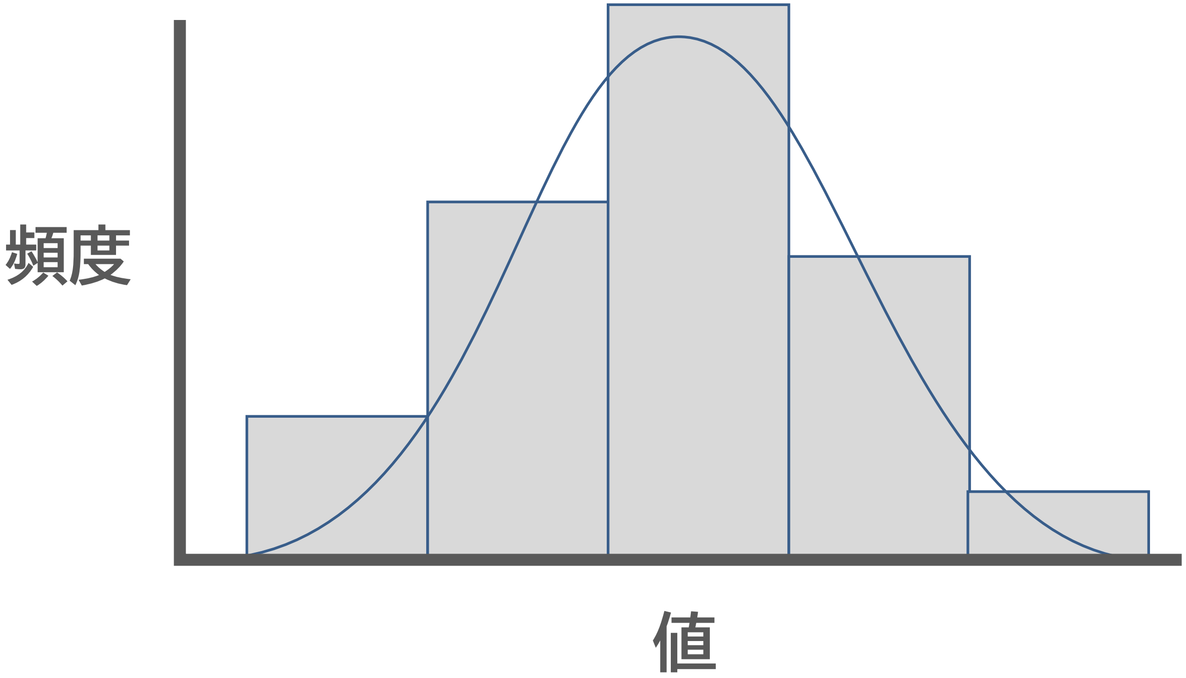
分散
標準偏差

ばらつきを表す値

統計量（基本統計量、基礎統計量とも）

分布

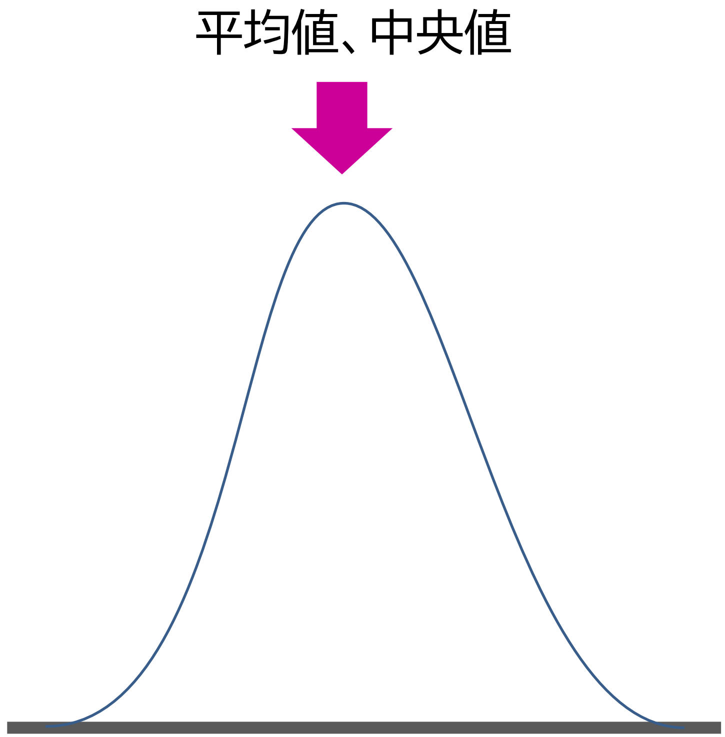
データの散らばり具合を表したもの



ヒストグラム（頻度分布図）

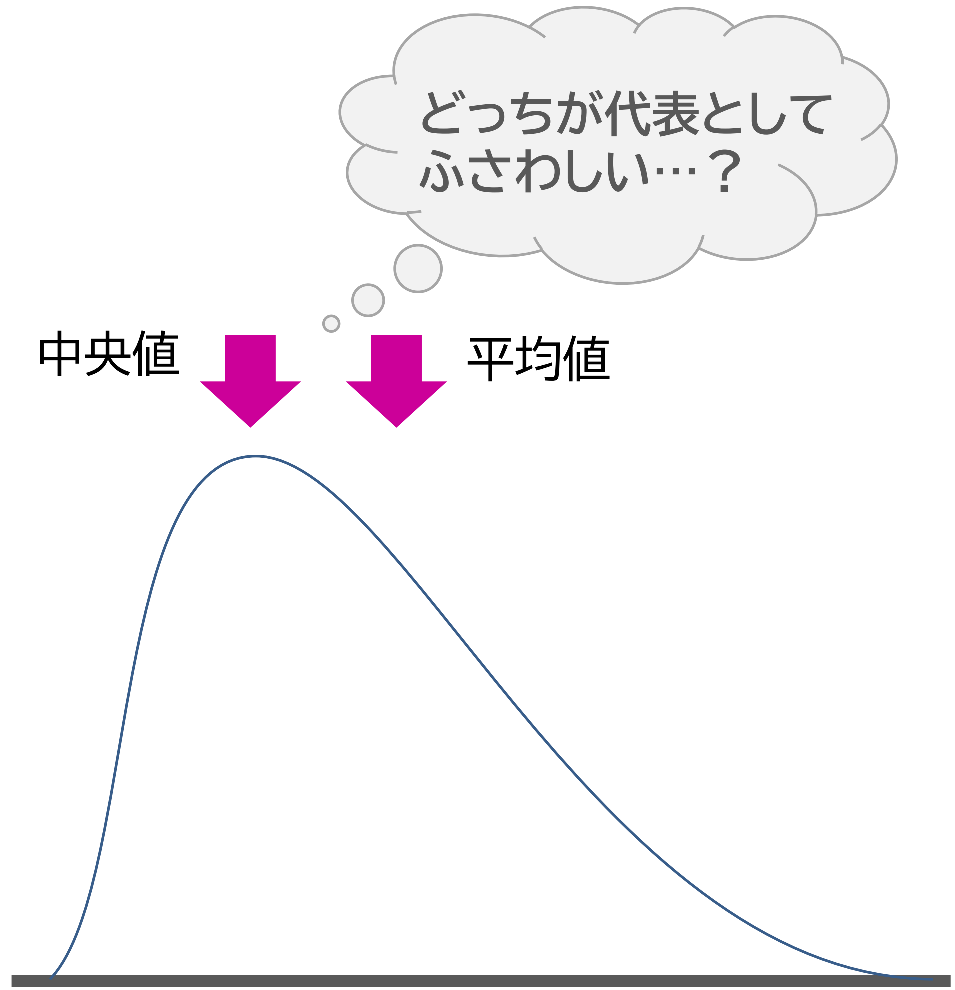
分布を使ったイメージ

データの中心



偏りのないデータ

- 身長の分布など

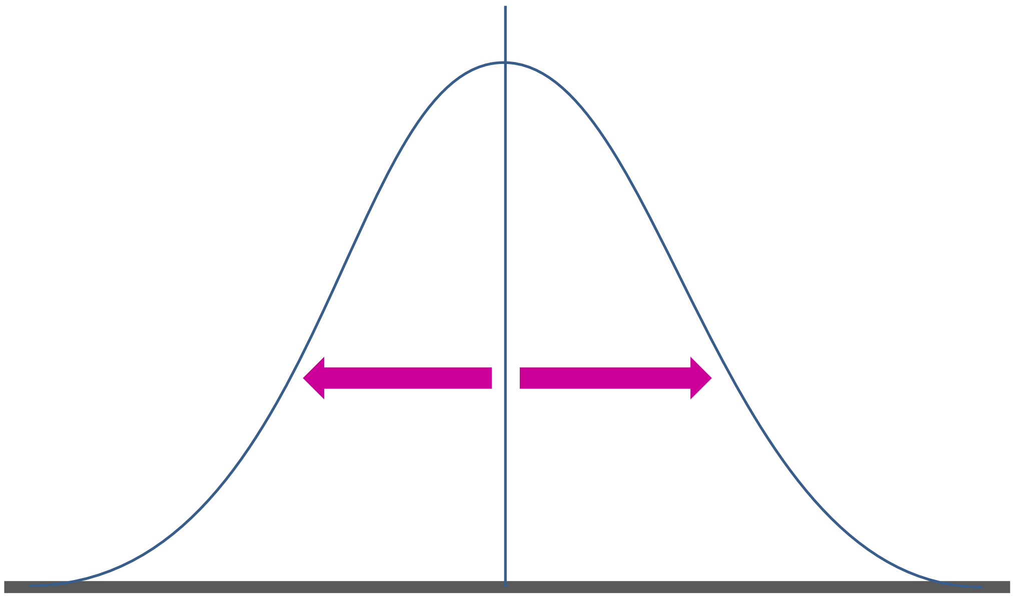


偏っているデータ

- 体重の分布
- 所得の分布など

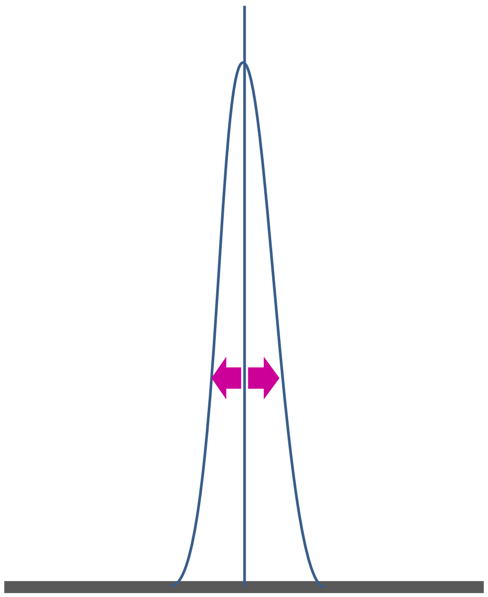
分布を使ったイメージ

ばらつき



ばらつき大きい

中心からの差が
全体的に大きい



ばらつき少ない

中心からの差が全
体的に小さい

計算方法

平均値

- 合計を計算
- 要素数で割る

計算方法

中央値

小さい順(大きい順)にならべて、
真ん中の値を取る

- 要素が奇数の場合、真ん中の値を採用
- 要素が偶数の場合、真ん中の2要素の平均値を計算

計算方法

分散、標準偏差

ばらつき

＝

平均値からの**ずれ**の大きさ

中央値ではなく

計算方法

分散

- 平均値を計算
- 各要素の値-平均値を計算
- その値を2乗
- その平均値を計算

分散

②要素iと平均値の差

①平均値

⑤要素数nで
割って平均
にする

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

③その2乗

④その全要素（iが1からnまで）の合計

分散 …2乗された値

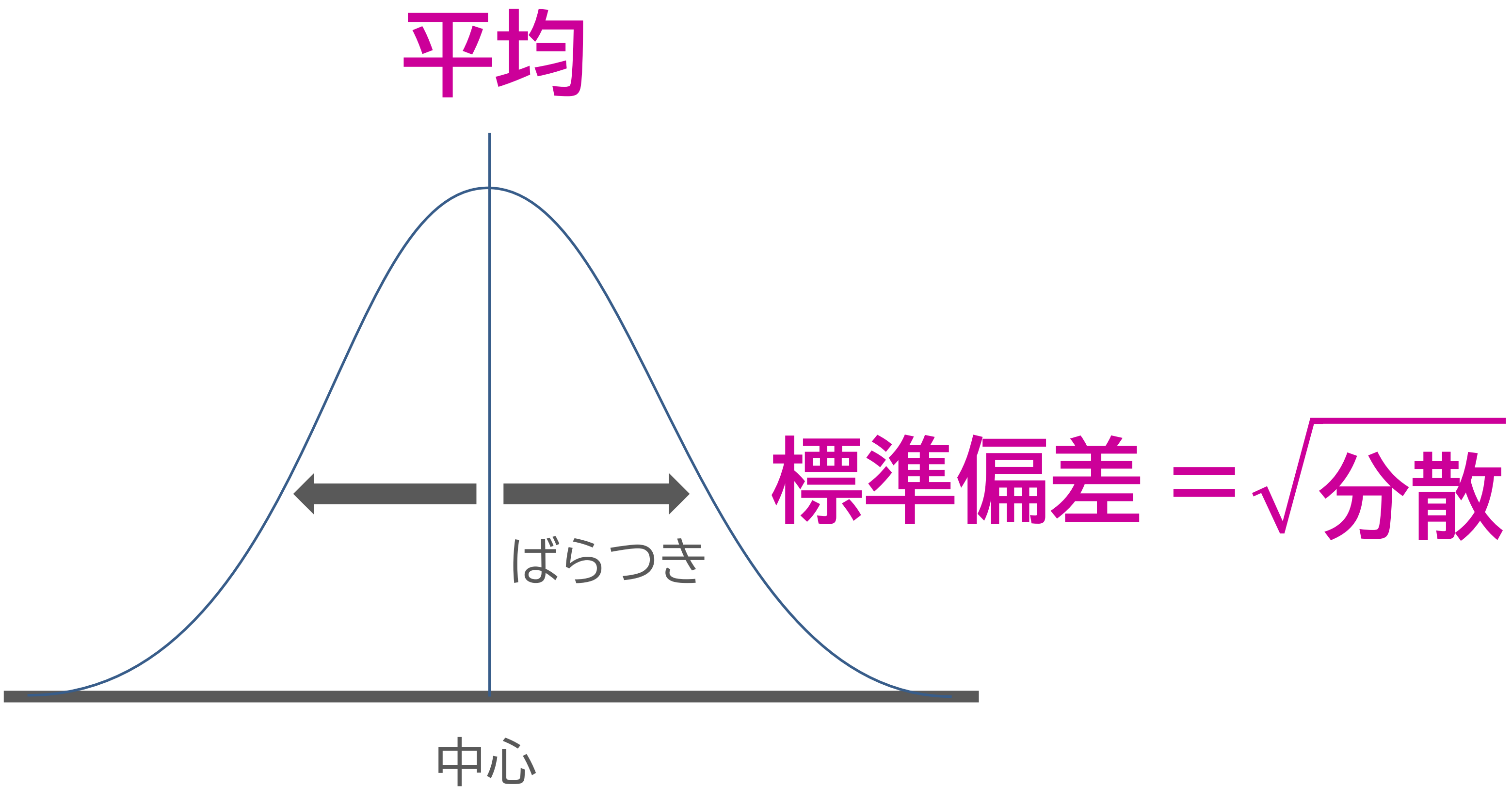
計測した値と単位を
そろえるため、
平方根を計算

標準偏差

目的:
このクラスの人の身長はどのくらい？

	平均		標準偏差
男性		±	
女性		±	

分布を使ったイメージ



もっと広い
世界が知りたい

目的:

このクラスの人の身長は
どのくらい？

目的:

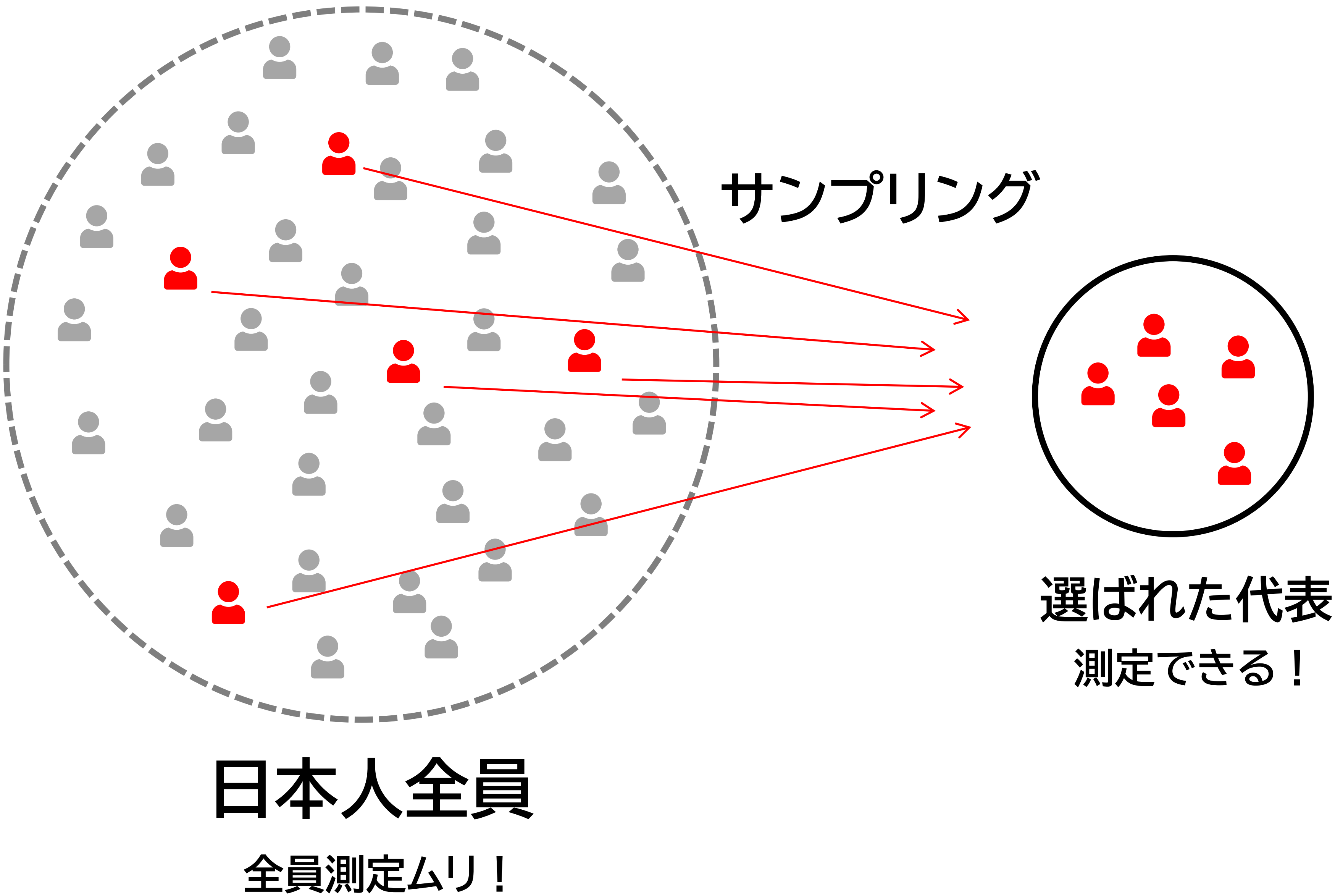
日本人の身長はどのくらい？

全員の身長を測定して計算する

✓ 現実的ではない
✓ コストもかかる

何名かを<u>抜き取り</u>調査する

サンプリング（抽出）


サンプリング
選ばれた代表
測定できる！
日本人全員
全員測定ムリ！

サンプリング

偏りなくランダムに選ぶことが原則

ランダムサンプリング
（無作為抽出）

サンプリングされた要素

標本
（サンプル）

今回の目的の場合、
サンプリングされた人のこと

サンプリング前の要素全体

母集団　＝解析の対象

今回の目的の場合、
日本人全員のこと

標本の数が多いほど、正確になる！

目的:

日本人の身長はどのくらい?

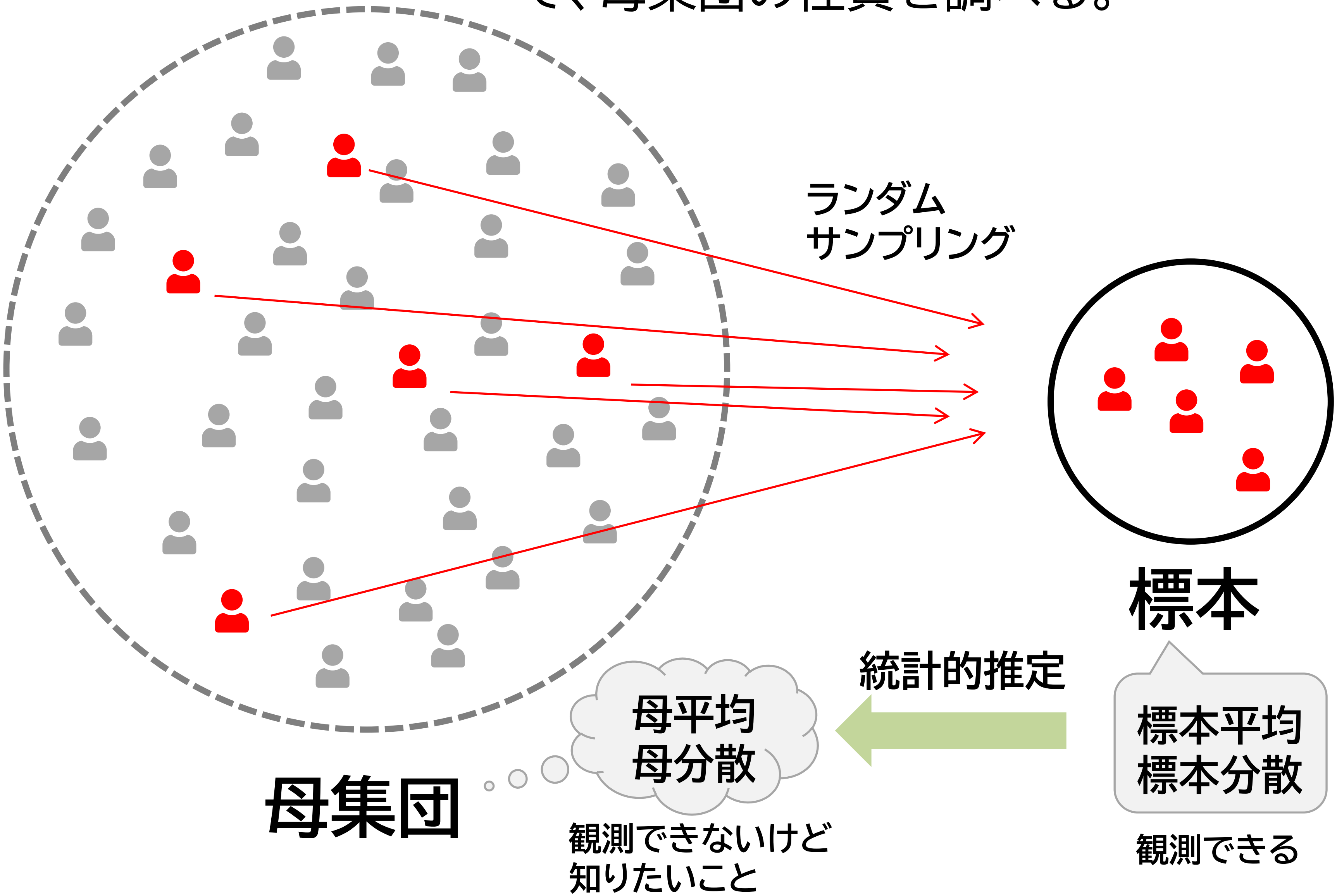
限られた標本を使って
母集団(日本人全体)の

- 推定の平均値や
- 推定のばらつき

を計算する、という問題

統計的推定

母集団が大きい、あるいは無限で、直接観測できないとき、標本を観測することで、母集団の性質を調べる。



母平均 μ ← 標本平均 $\bar{x}$

一致が期待できる

母分散 σ^2 ~~←~~ 標本分散 s^2

実は一致が期待できない!!

一致が期待できるのは、母集団の全標本を観測できる場合(全数検査)だけ

不偏(標本)分散 v^2

一致が期待できる

真の値から外れていないことを、
不偏性があると言うので

標本分散

②要素iと平均値の差

①標本平均

⑤要素数nで割って平均にする

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

③その2乗

④その全要素（iが1からnまで）の合計

不偏（標本）分散

⑤n-1で割る

$$\frac{1}{\boldsymbol{n-1}}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

n-1で割る？

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

- 標本の数nが母集団の数N（大きな数）に近づくと、母分散に近くなる

➡ **母分散の推定に使える**

- 自由度を表している

自由度 ＝ 互いに影響を与えない（独立した）値の個数

上の式で、一つの観測値x（i=a）は他と完全に独立ではなく、それ以外の（n-1）個の独立した観測値と平均値$\bar{x}$によって求められる。

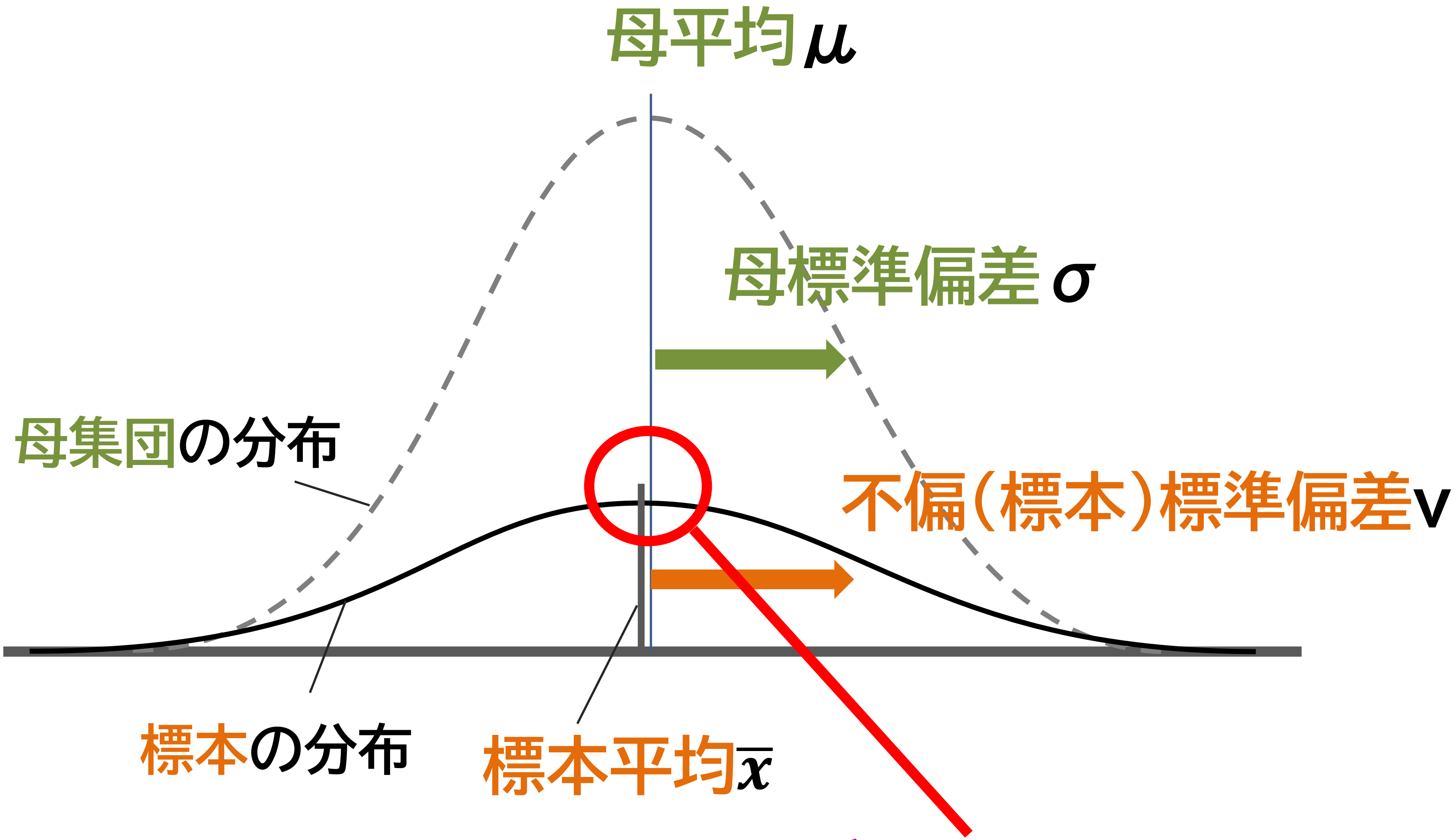
用語より、
n-1で割っているか
どうかに注目

書籍によって、標本分散s^2を不偏標本分散（不偏分散）のこととして記述しているものもあります。「（不偏）標本分散」と記述されることもあります。標本を考える時点で、そもそも母集団の推定を前提としていることが多いためです。

nで割っていたら、**観測値**の話
n-1で割っていたら、**推定値**の話

です

イメージ

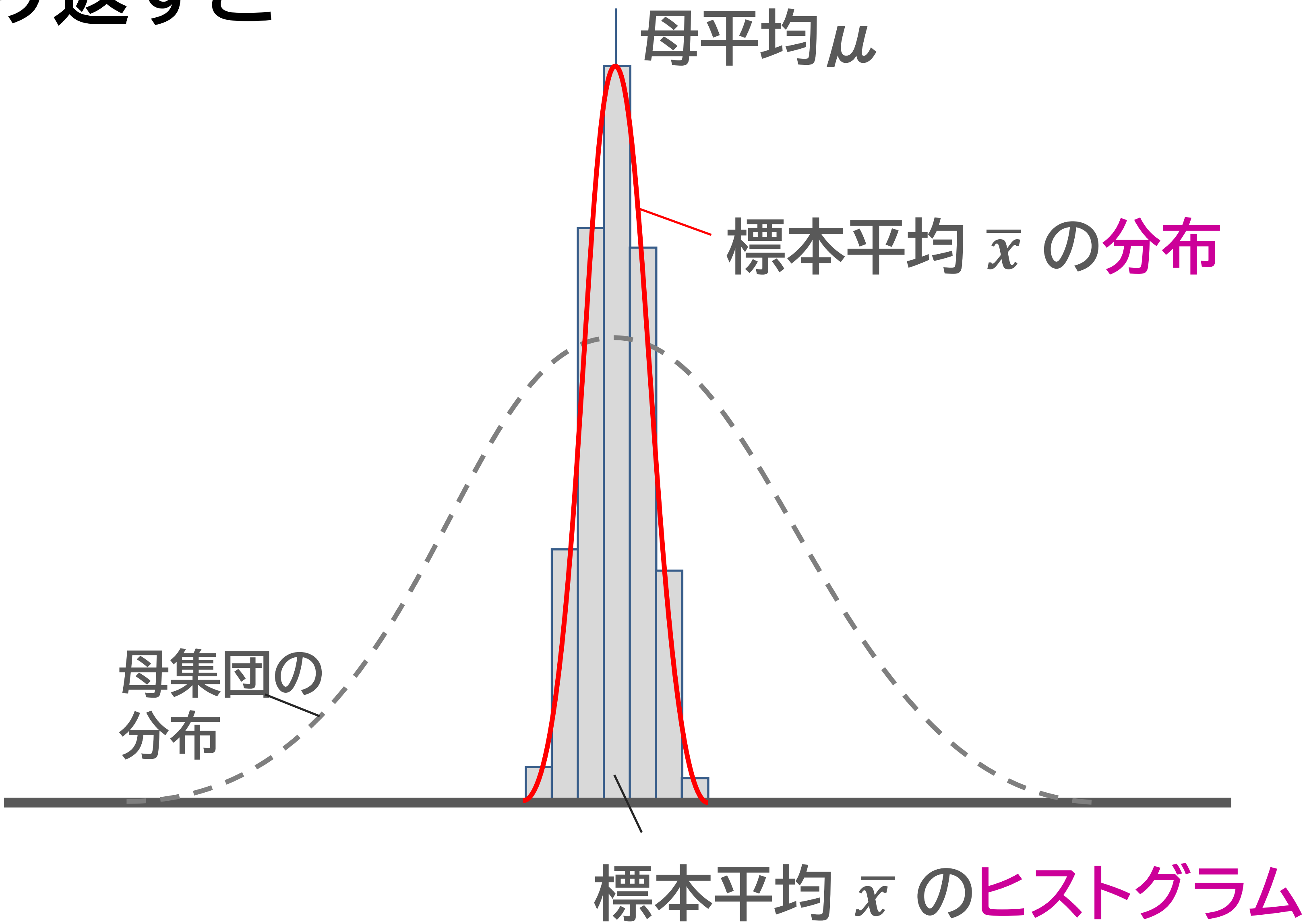


きっと**ズレ**が生じている

誤差

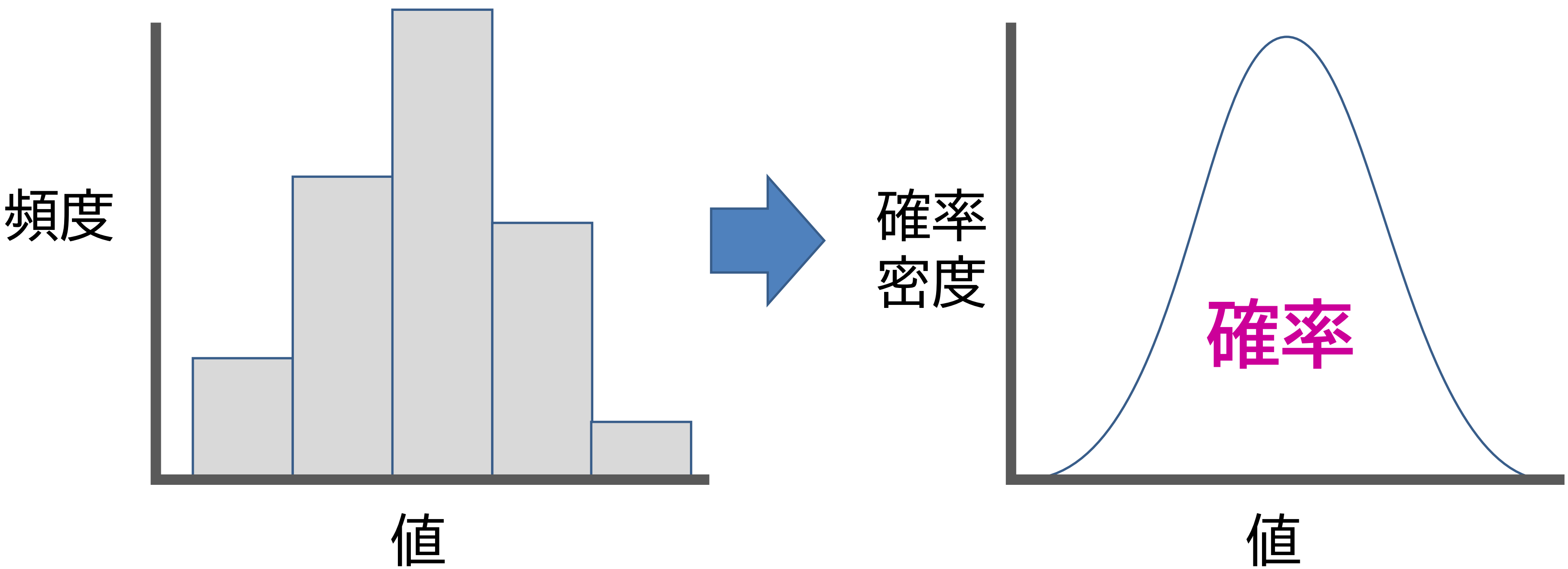
- サンプリング誤差
- 測定誤差

サンプリングして標本平均$\bar{x}$を算出して、を繰り返すと…



分布

データの散らばり具合を表したもの



ヒストグラム

観測結果を表したもの

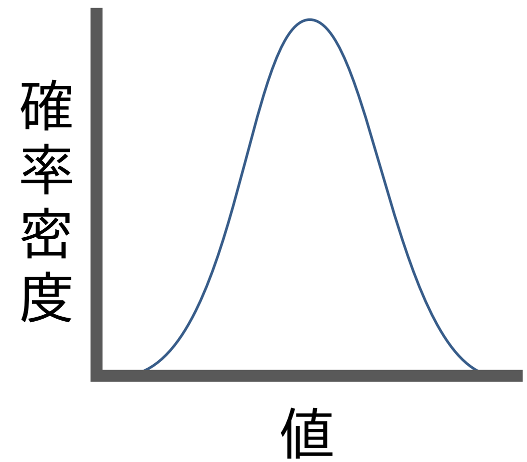
確率密度関数

事象の起こる確率を表したもの

代表的な分布

正規分布（ガウス分布）

- 平均値が中心で、
- 平均値に近いものが多く、
- 左右に均等な釣り鐘状の分布



均等な確率で生じたばらつきの場合にとる分布

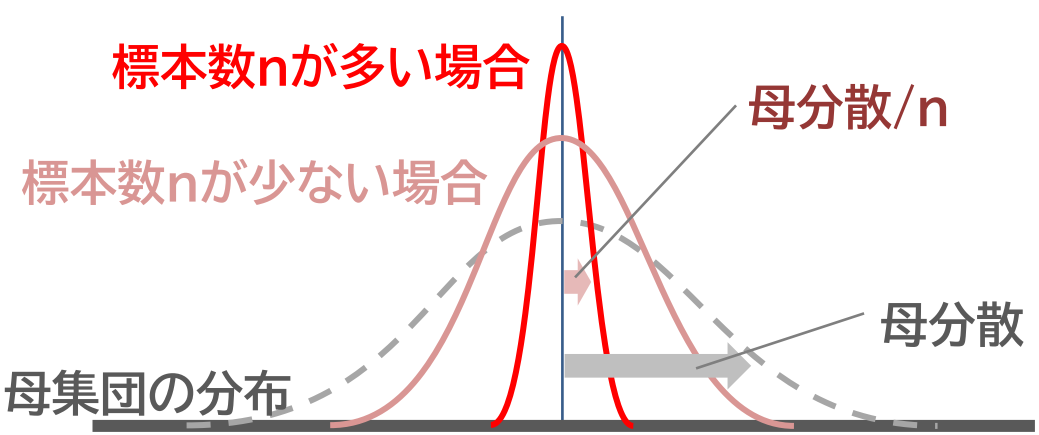
✓ 身長の分布
✓ 測定誤差の分布
✓ 自然界で起こるゆらぎ　　など

標本平均$\bar{x}$の分布

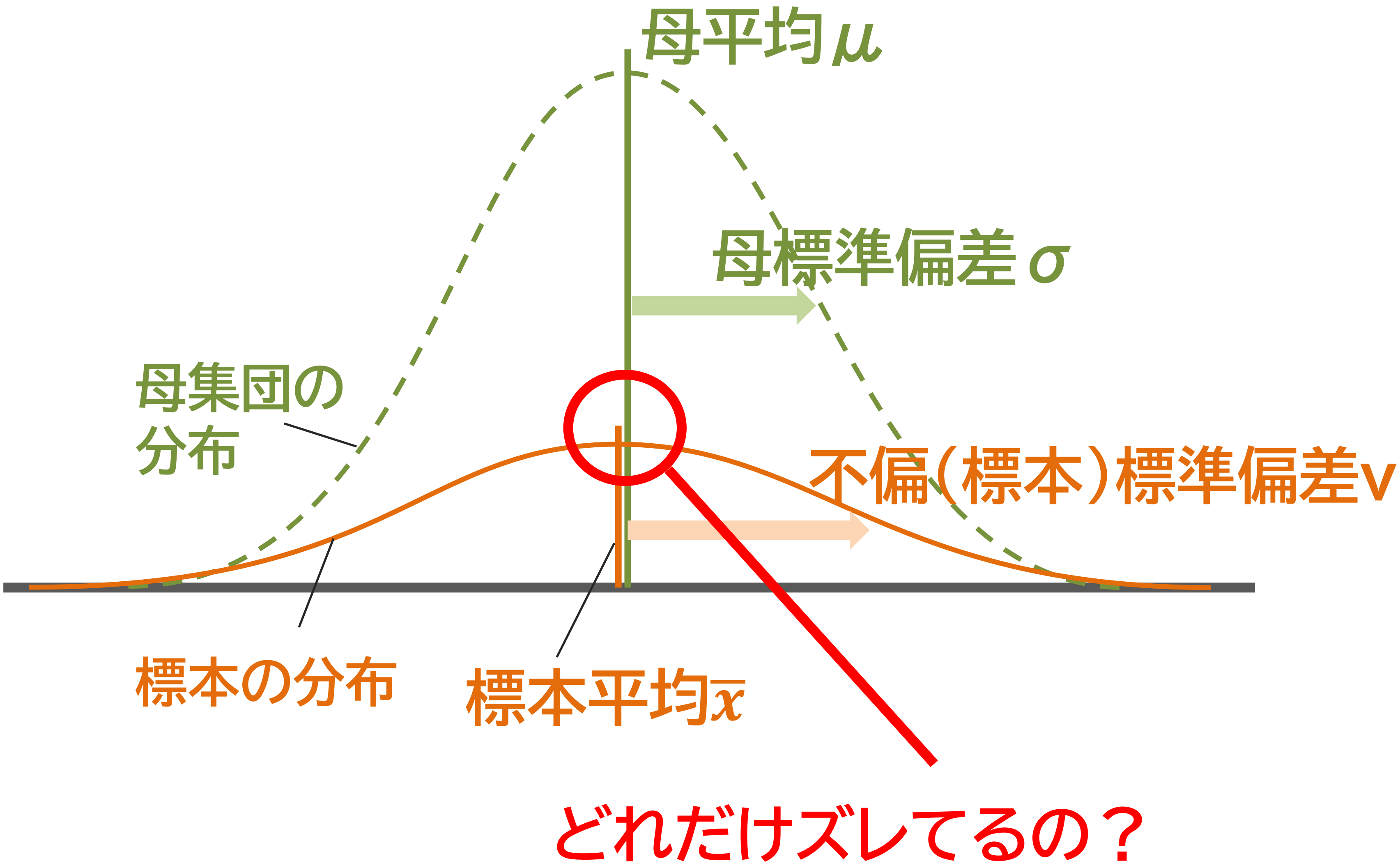
- 正規分布に従う
- 分散は、標本数 n が大きいほど、小さくなる

n=母集団数N なら、全数検査なので、母平均μとのずれはゼロになる。
n=1 なら、母集団のうち一つずつを測定するのと同じなので、分散も同じ。

- 分散は、**母分散σ^2の1/n**になる



中心極限定理



ずれの大きさを、標本数nで示せる!!

標準誤差

母平均μの推定値のばらつきを表したもの

＝　標本平均$\bar{x}$の分布の標準偏差

分散の平方根

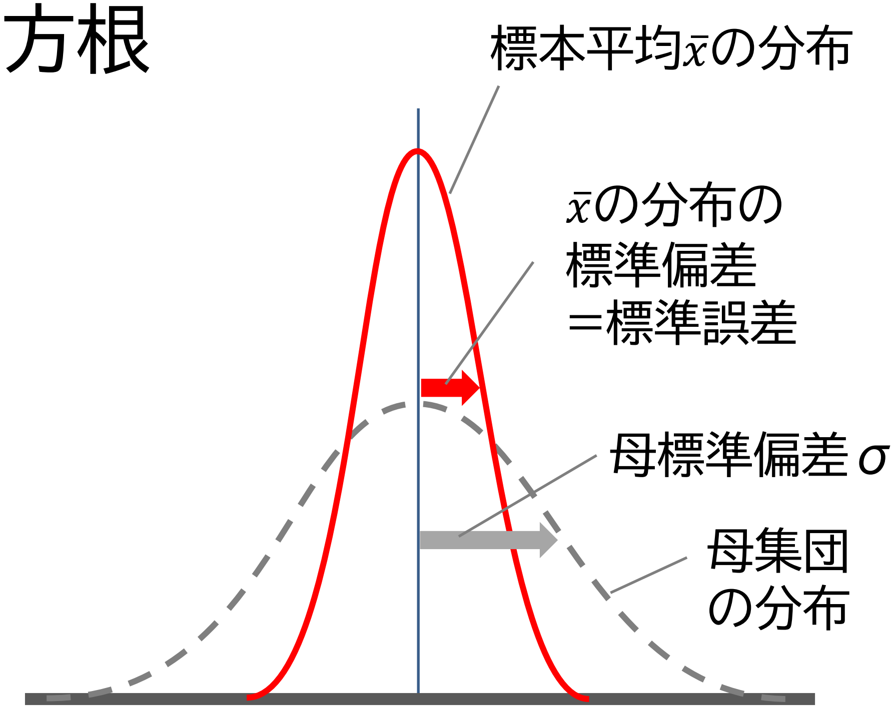
中心極限定理から

＝　母分散σ^2の1/n、の平方根

$$= \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

$$= \frac{1}{\sqrt{n}} \times \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

不偏標本標準偏差



標準偏差 と 標準誤差

論文などでよく見る図

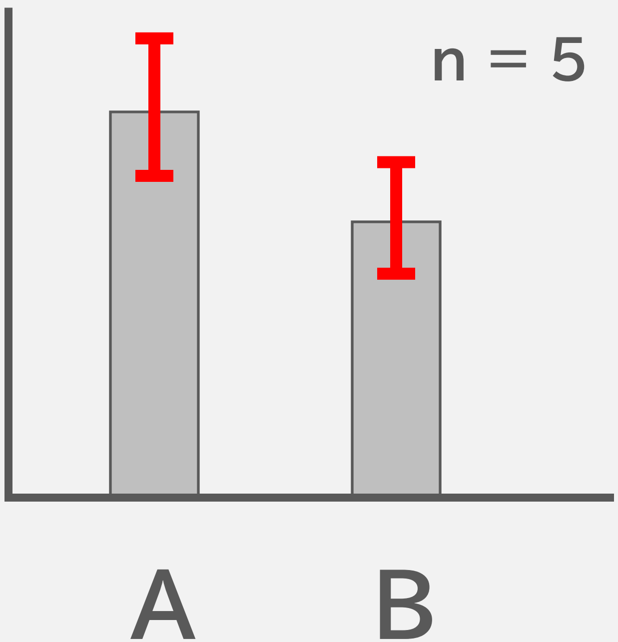


図1 A群とB群の＊＊の違い
それぞれ5個体を測定した。エラーバーは標準偏差を表す

エラーバーが**標準偏差**

↓

測定した標本自体の平均値を論じている

エラーバーが**標準誤差**

↓

測定した標本から推定される母集団の平均値について論じている

【ここに注意！】

標準誤差は標準偏差の$1/\sqrt{n}$なので、エラーバーは短くなり、より明確な差がありそうな見栄えになります。標準誤差を示すことが適当なのかどうかを、正しく判断しながらデータを解釈しましょう。

計算してみよう

このクラスの身長データからいくつかのデータを抜き出し、クラスの身長の平均値を推定してみる